PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-256253
(43)Date of publication of application : 21.09.2001

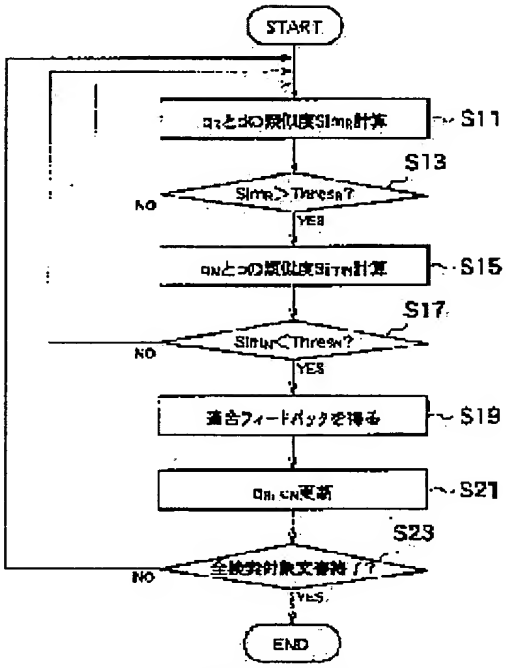
(51)Int.Cl. G06F 17/30

(21)Application number : 2000-069477
(22)Date of filing : 13.03.2000

(71)Applicant : KDDI CORP
(72)Inventor : HOASHI KEIICHIRO
INOUE NAOKI
MATSUMOTO KAZUNORI
HASHIMOTO KAZUO

(54) METHOD AND DEVICE FOR FILTERING DOCUMENT

(57)Abstract:
PROBLEM TO BE SOLVED: To provide a method and a device for filtering document, with which the selection of an unsuited document can be avoided by generating an unsuited profile expressing the features of the erroneously selected unsuited document, and not selecting the document of high similarity to the unsuited profile.
SOLUTION: Similarity SimR between a suited profile qR and a retrieval object document (d) is calculated (step S11), this similarity SimR is compared with a threshold ThresR and when SimR is greater than ThresR, similarity SimN between an unsuited profile qN and the retrieval object document (d) is calculated (step S15). When SimN is greater than ThresN, the retrieval object document is regarded as an unsuited document and excluded. When SimN is smaller, this document is regarded as a suited document and selected, suited feedback is provided and qR and qN are updated (step S21).



LEGAL STATUS

[Date of request for examination] 10.02.2003

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998;2003 Japan Patent Office

(19)日本国特許庁（ＪＰ）　(12) 公開特許公報（Ａ）　(11)特許出願公開番号
特開2001－256253
（P2001－256253A）
(43)公開日　平成13年９月21日(2001.9.21)

(51)Int.Cl.⁷	識別記号	ＦＩ	ﾃｰﾏｺｰﾄﾞ(参考)
Ｇ０６Ｆ　17/30	３４０ １７０	Ｇ０６Ｆ　17/30	３４０Ａ　５Ｂ０７５ １７０Ａ

審査請求　未請求　請求項の数４　ＯＬ　（全 14 頁）

(21)出願番号　特願2000－69477(P2000－69477)

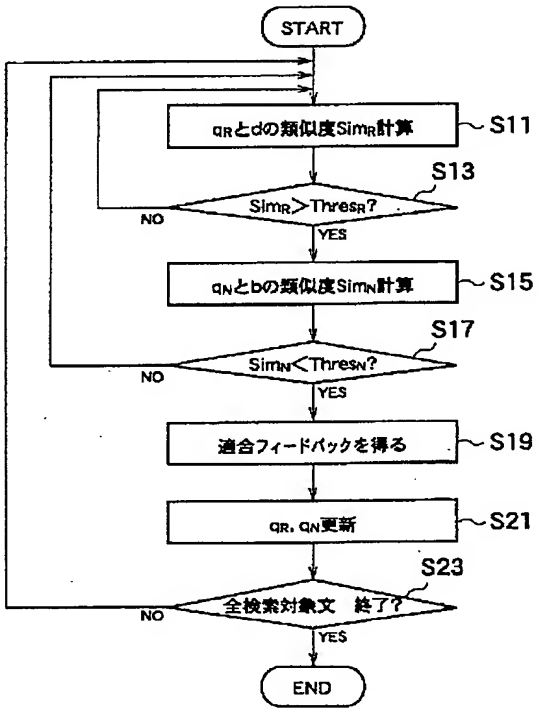
(22)出願日　平成12年３月13日(2000.3.13)

(71)出願人　000208891
ケイディーディーアイ株式会社
東京都新宿区西新宿二丁目３番２号
(72)発明者　帆足　啓一郎
埼玉県上福岡市大原２－１－15　株式会社ケイディディ研究所内
(72)発明者　井ノ上　直己
埼玉県上福岡市大原２－１－15　株式会社ケイディディ研究所内
(74)代理人　100083806
弁理士　三好　秀和　（外３名）

最終頁に続く

(54)【発明の名称】　文書フィルタリング方法および装置

(57)【要約】
【課題】　誤って選択された非適合文書の特徴を表す非適合プロファイルを作成し、非適合プロファイルとの類似度が高い文書を選択しないということで非適合文書の選択を回避し得る文書フィルタリング方法および装置を提供する。
【解決手段】　適合プロファイルq_Rと検索対象文書ｄとの類似度Sim_Rを計算し（ステップＳ１１）、該類似度Sim_Rと閾値$Thres_R$と比較し、Sim_Rが$Thres_R$より大きい場合、非適合プロファイルq_Nと検索対象文書ｄとの類似度Sim_Nを計算し（ステップＳ１５）、Sim_Nが$Thres_N$より大きい場合、検索対象文書を非適合文書と見なして除外し、小さい場合、適合文書と見なして選択し、適合フィードバックを得て、q_R，q_Nを更新する（ステップＳ２１）。



【特許請求の範囲】
【請求項１】　ユーザの要求を表す適合プロファイルに適合する文書を検索対象文書の中から抽出して出力する文書フィルタリング方法であって、
前記適合プロファイルに対して類似度が高いが、ユーザの要求に適合しないと判定された非適合文書を特徴づけるプロファイルを非適合プロファイルとして作成し、
前記適合プロファイルと検索対象文書との類似度を算出し、
この算出した類似度を所定の適合用閾値と比較し、
該類似度が所定の適合用閾値より大きい場合、前記検索対象文書と前記非適合プロファイルとの類似度を算出し、
この非適合プロファイルとの類似度を所定の非適合用閾値と比較し、
該類似度が所定の非適合用閾値より大きい場合、該検索対象文書を非適合文書と見なして除外し、小さい場合、適合文書と見なして選択し、
前記非適合文書および適合文書の適合フィードバックを行い、適合プロファイルおよび非適合プロファイルを更新することを特徴とする文書フィルタリング方法。
【請求項２】　前記非適合プロファイルの更新は、
選択文書に出現する単語から単語寄与度に基づき選択文書の特徴を表す単語を抽出し、
この抽出された単語の寄与度を算出し、
この単語寄与度に重みをかけて、単語に対するスコアを算出し、
前記選択文書が適合文書である場合には、前記スコアを非適合プロファイルから減算し、前記選択文書が非適合文書である場合には、前記スコアを非適合プロファイルに加算して更新することを特徴とする請求項１記載の文書フィルタリング方法。
【請求項３】　ユーザの要求を表す適合プロファイルに適合する文書を検索対象文書の中から抽出して出力する文書フィルタリング装置であって、
前記適合プロファイルに対して類似度が高いが、ユーザの要求に適合しないと判定された非適合文書を特徴づけるプロファイルを非適合プロファイルとして作成する非適合プロファイル作成手段と、
前記適合プロファイルと検索対象文書との類似度を算出する適合プロファイル類似度算出手段と、
この算出した類似度を所定の適合用閾値と比較し、該類似度が所定の適合用閾値より大きい場合、前記検索対象文書と前記非適合プロファイルとの類似度を算出する非適合プロファイル類似度算出手段と、
この非適合プロファイルとの類似度を所定の非適合用閾値と比較し、該類似度が所定の非適合用閾値より大きい場合、該検索対象文書を非適合文書と見なして除外し、小さい場合、適合文書と見なして選択する選択手段と、
前記非適合文書および適合文書の適合フィードバックを行い、適合プロファイルおよび非適合プロファイルを更新する更新手段とを有することを特徴とする文書フィルタリング装置。
【請求項４】　前記更新手段は、
選択文書に出現する単語から単語寄与度に基づき選択文書の特徴を表す単語を抽出する単語抽出手段と、
この抽出された単語の寄与度を算出する単語寄与度算出手段と、
この単語寄与度に重みをかけて、単語に対するスコアを算出するスコア算出手段と、
前記選択文書が適合文書である場合には、前記スコアを非適合プロファイルから減算し、前記選択文書が非適合文書である場合には、前記スコアを非適合プロファイルに加算して更新するスコア加減算手段とを有することを特徴とする請求項３記載の文書フィルタリング装置。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、ユーザの要求を表す適合プロファイルに適合する文書を検索対象文書の中から抽出して出力する文書フィルタリング方法および装置に関し、更に詳しくは、適合プロファイルとの類似度は高いが、ユーザの要求に適合しないと判定された非適合文書を特徴づけるプロファイルを非適合プロファイルとして作成し、この非適合プロファイルに対して類似度が高い文書を選択しないように削除して文書のフィルタリングを行う文書フィルタリング方法および装置に関する。
【０００２】
【従来の技術】この種の文書フィルタリング技術は、例えばメールによるニュース情報配信サービスなどのように大量のテキスト情報の流れの中からユーザの嗜好に合致した情報のみを抽出して、ユーザに提供するのに有効である。すなわち、文書フィルタリングとは、次々に流れてくる検索対象文書の中からユーザの要求を満たしている文書のみを取得し、ユーザに提供するタスクである。
【０００３】このような文書フィルタリングにおいて、ユーザの要求はプロファイルとしてフィルタリングシステム内で表されている。そして、フィルタリングシステムは、順次流れてくる検索対象文書１つ１つについて、このプロファイルを満たしているか否かを判断し、要求を満たしている文書のみをユーザに提示する。ユーザは、提示された文書に対して実際に要求を満たしているか否かを判断し、その判断をフィルタリングシステムにフィードバックする。多くの場合、フィルタリングシステムはユーザからのフィードバックを基にプロファイルを更新することによりフィルタリングの精度向上を図っている。
【０００４】フィルタリングシステムには多くの場合情報検索で使用される技術が適用されている。システムに

入力される各文書やプロファイルはベクトル空間モデルなどに基づいてシステム内で表現され、各文書がプロファイルを満たしているか否かの判断基準として、プロファイルと文書との間の類似度が使用されることが多い。また、文書フィルタリングのプロファイル更新には情報検索における検索式拡張手法を適用する手法が多く用いられている。すなわち、ユーザからの適合フィードバック情報に基づき、選択された文書から抽出された情報をプロファイルに追加して、プロファイルを更新することにより、プロファイルを精緻化させる処理を行うのである。

【０００５】このようなプロファイル更新を利用した従来の文書フィルタリング方法の処理手順について図５に示すフローチャートを参照して説明する。

【０００６】この文書フィルタリング方法では、まずユーザの要求であるプロファイルｑに類似する文書を検索対象文書ｄの中から検索すべくプロファイルｑと検索対象文書ｄの類似度を計算する（ステップＳ７１）。この計算したプロファイルｑと検索対象文書ｄとの類似度が所定の閾値より大きいか否かを判定する（ステップＳ７３）。該類似度が所定の閾値よりも大きくない場合には、最初のステップＳ７１に戻り、次の検索対象文書について同じ処理を繰り返すが、類似度が所定の閾値よりも大きい場合には、適合フィードバックを得て（ステップＳ７５）、プロファイルｑを更新する（ステップＳ７７）という処理を全ての検索対象文書について行って、処理を終了する（ステップＳ７９）。

【０００７】このような文書フィルタリング方法では、一般的にプロファイルとの類似度が閾値を上回る文書を選択してユーザに提示するという流れで文書のフィルタリングを行っているが、このようなフィルタリング手法では後述するように適切な閾値を設定することが難しく、多くの適合文書を選択するために閾値を低く設定すると、誤って選択される非適合文書の数が大きく増大し、また逆に非適合文書の誤り選択を減少させるために閾値を高く設定すると、多くの適合文書を見逃してしまうことになる。

【０００８】このような文書フィルタリング方法におけるプロファイルの更新手法には上述したように情報検索で使用される検索式拡張技術が適用されることが多いが、次に情報検索において高い精度が得られる単語寄与度に基づく検索式拡張手法を適用したプロファイル更新手法について説明する。

【０００９】まず、単語寄与度に基づく検索式拡張手法について説明する。単語寄与度とは、文書間の類似度における各単語の影響を数値化した尺度である。ある入力文ｑと検索対象文書ｄとの間の類似度における単語w_iの単語寄与度を式（１）によって定義する。

【００１０】

【数１】

$$Cont(w_i, q, d) = Sim(q, d) - Sim(q'(w_i), d'(w_i)) \quad \cdots (1)$$

ただし、Ｓｉｍ(ｑ，ｄ）は、ｑ，ｄ間の類似度を表し、$q'(w_i)$はｑから単語w_iを除いた入力文を、$d'(w_i)$はｄから単語w_iを除いた文書を表すとする。

【００１１】すなわち、単語寄与度Ｃｏｎｔ(w_i，ｑ，ｄ）とは、ｑとｄとの類似度と単語w_iが存在しない場合のｑとｄとの類似度との差である。従って、ｑとｄに出現する全ての単語のうち、類似度を向上させる単語の寄与度は正であり、逆に類似度を下げる単語の寄与度は負である。

【００１２】また、文献「帆足、松本、井ノ上、橋本：文書間の類似度における単語寄与度を利用した検索式拡張手法、情報処理学会論文誌：データベース、Vol.40,No.SIG8(TOD4),pp.63-73,1999」によれば、出現単語の多くの寄与度は０に近く、類似度に有意な影響を与えている単語は少ない。そのうち、寄与度が大きく正の値を持つ単語は、入力文と検索対象文書の両方に存在した単語である。一方、大きな負の値の寄与度を持つ単語は一方の文書にのみ存在し、かつ、その文書の特徴を顕著に表す単語であると考えられる。そこで、単語寄与度に基づいた検索式拡張手法では以下のように検索式の拡張を行っている。

【００１３】まず、入力文ｑと適合している文書群

$$D_{rel}(q) = \{d_1, \cdots, d_{Num}\} \quad \cdots (2)$$

中の各文書に出現する全ての単語の寄与度を求め、各類似文書から単語寄与度の低い単語をＮ個抽出する。次に抽出された各単語の寄与度の総和に重みｗｇｔをかけ、これを単語ｗに対するスコアとする。単語ｗの入力文ｑと文書ｄの類似度に対する寄与度をＣｏｎｔ(w_i，ｑ，ｄ）とすると、単語ｗのスコアＳｃｏｒｅ(ｗ）は式（３）によって表される。

【００１４】

【数２】

$$Score(w) = wgt \times \sum_{d \in D_{rel}(q)} Cont(w, q, d) \quad \cdots (3)$$

次に、抽出された単語のうち元の検索式に含まれていない単語を検索式に加えることで検索式拡張を実現する。

【００１５】ある単語ｗを入力文のベクトルに加える際には、式（３）で計算されたスコアＳｃｏｒｅ(ｗ）を単語ｗが入力文に出現する頻度（単語出現頻度　ｔｆ）と見なし、入力文のベクトル内で単語ｗを表す要素の値

を計算する。ベクトルの各要素がＴＦ*ＩＤＦによって計算されている場合、Ｓｃｏｒｅ(ｗ)をｔｆとすることでＴＦを算出し、更に単語ｗのＩＤＦをかけ、その結果得られたＴＦ*ＩＤＦ値を入力文のベクトルの単語ｗの要素に入れることにより、検索式拡張を行う。

【００１６】次に、単語寄与度に基づくプロファイル更新手法について説明する。

【００１７】単語寄与度による検索式拡張では、初期検索の結果に対するフィードバックにより得られた適合文書集合中の各文書から寄与度に基づいて抽出された単語の寄与度の総和に重みを掛けることで、各単語に対するスコアを求めた。ここでは、フィルタリング中に選択された文書１つ毎に単語寄与度に基づき文書中から単語を抽出し、その情報を直前のプロファイルに加えることで、随時プロファイルを更新する。

【００１８】まず、選択された文書が適合文書の場合には、抽出した単語w_iに対するスコア$Score_{rel}(w_i)$を式（４）により算出し、選択した文書が非適合文書中の場合には、抽出した単語w_iのスコア$Score_{nrel}(w_i)$を式（５）により算出する。

【００１９】

【数３】

$$Score_{rel}(w_i) = wgt_{relR} \times Cont(w_i, q, d) \quad \cdots (4)$$

$$Score_{nrel}(w_i) = wgt_{nrelR} \times Cont(w_i, q, d) \quad \cdots (5)$$

上記の式によって求めた各単語のスコアを単語出現頻度ｔｆとして扱い、ＴＦ*ＩＤＦ法により各単語の重みを算出する。そして、抽出された単語が適合文書中の単語の場合はその単語と重みをプロファイルに加え、非適合文書中の単語の場合は単語と重みをもとのプロファイルから引く。すなわち、適合文書から選択された各単語の要素が元のプロファイルに加えられ、非適合文書から選択された各単語の要素が元のプロファイルから引かれるということである。なお、この処理により負の重みを持った単語は、類似度計算に使用されない。

【００２０】以上の処理により、プロファイルを表すベクトルの、値を持たなかった次元が正の値を持つようになり、プロファイルの情報が拡張されることになる。また、適合文書と非適合文書に共起する単語の重みが抑制され、適合文書のみに出現する単語の重みが強調される。

【００２１】ここでは検索対象文書とプロファイルの両方をベクトル空間モデルを用いて表し、両者間の類似度を計算することで各文書に対するフィルタリングを実現した。

【００２２】ベクトル空間モデルで表現する際、各文書およびプロファイルを表すベクトルの各要素の重みはＴＦ*ＩＤＦ法により算出する。ここでは、最も有効な情報検索システムの１つであるＳＭＡＲＴにおいて使用されているアルゴリズムに基づきＴＦおよびＩＤＦの計算式を使用した。本実験で使用したＴＦおよびＩＤＦの計算式を式（６）、式（７）に示す。

【００２３】

【数４】

TF factor

$$\log(1 + tf_{ij}) \quad \cdots (6)$$

IDF factor

$$\log\left(\frac{M}{df_j}\right) \quad \cdots (7)$$

ただし、tf_{ij}は文書d_i内の単語w_jの出現頻度、df_jは単語w_jが出現する文書数、Ｍは語彙作成時に使用された文書集合に含まれる文書数とする。

【００２４】また、類似度は式（８）で定義されるプロファイルと検索対象文書のベクトルのコサイン値を取ることにより正規化した値として求めた。

【００２５】

【数５】

$$\cos(\vec{q}, \vec{d}) = \frac{\vec{q} \cdot \vec{d}}{|\vec{q}||\vec{d}|} \quad \cdots (8)$$

ただし、$\vec{q}$, $\vec{d}$はそれぞれ入力文書と検索対象文書を表すベクトルとし、$|\vec{d}|$は$\vec{d}$のユークリッド長とする。

本プロファイル更新手法では、式（４），（５）によって求めた各単語のスコアを単語出現頻度ｔｆとして扱い、式（６），（７）によりＴＦおよびＩＤＦを計算する。従って、各単語のＴＦ*ＩＤＦ値は式（９），（１０）により算出される。

【００２６】

【数６】

$$Value_{rel}(w_i) = log(1+Score_{rel}(w_i)) \times log\left(\frac{M}{df_i}\right) \quad \cdots (9)$$

$$Value_{nrel}(w_i) = log(1+Score_{nrel}(w_i)) \times log\left(\frac{M}{df_i}\right) \quad \cdots (10)$$

ただし、df_iは単語w_iが出現する文書数、Ｍはｄｆのリストの作成に使用された文書数とする。

【００２７】また、プロファイルｑおよび文書ｄをベクトル空間モデルによってそれぞれ式（１１），（１２）のように表すとする。

【００２８】

【数７】

$$\vec{q} = (q_1, \cdots, q_n) \quad \cdots (11)$$

$$\vec{d} = (d_1, \cdots, d_n) \quad \cdots (12)$$

ただし、q_1，…，q_nはプロファイル中の単語の重みを、d_1，…，d_nは各文書中の単語の重みを表し、ｎはベクトルの次元数を表す。

【００２９】更新後のプロファイルを式（１３）で表すとすると、抽出された各単語w_iについて、適合文書中の単語の場合は式（１４）により表され、非適合文書中の単語の場合は式（１５）により表される。

【００３０】

【数８】

$$\vec{q}_{new} = (q_1', \cdots, q_n') \quad \cdots (13)$$

$$q_i' = q_i + Value_{rel}(w_i) \quad \cdots (14)$$

$$q_i' = q_i - Value_{nrel}(w_i) \quad \cdots (15)$$

すなわち、適合文書から選択された各単語の要素が元のプロファイルの要素に加えられ、非適合文書から選択された各単語の要素が元のプロファイルの要素から引かれるということである。なお、この処理により負の重みを持った単語は、類似度計算に使用されない。

【００３１】図６は、上述したプロファイル更新手法の手順を示すフローチャートである。同図に示すように、ベクトル空間モデルによって表されたプロファイルｑおよび更新後のプロファイルq_{new}、および選択文書ｄに対して、まず選択文書ｄから単語集合Ｗを抽出し（ステップＳ８３）、選択文書ｄが適合文書であるか否かを判定する（ステップＳ８５）。選択文書ｄが適合文書である場合には、各単語に対するスコアを式（４）のように算出し（ステップＳ８７）、また（１４）で示したように各単語w_iのスコアをプロファイルｑに加える（ステップＳ８９）。また、選択文書ｄが適合文書でない場合には、各単語に対するスコアを式（５）のように算出し（ステップＳ８８）、また（１５）で示したように各単語w_iのスコアをプロファイルｑから引き（ステップＳ９１）、このような加減算したプロファイルｑを更新後のプロファイルq_{new}として設定する（ステップＳ９３）。

【００３２】上述したプロファイル更新手法に基づき、ＴＲＥＣ－８の Filtering Trackで用意されたデータを使用した評価実験を行った。図７に類似度の閾値を０．１に設定した場合に選択された適合文書および非適合文書のプロファイルとの類似度を示す。

【００３３】図７から閾値を大きく超える類似度を有す

る非適合文書は少ないものの、閾値近辺の類似度では適合文書と非適合文書が混在していることがわかる。

【００３４】

【発明が解決しようとする課題】上述したように、従来の文書フィルタリング方法では、閾値を大きく超える類似度を有する非適合文書は少ないものの、閾値近辺の類似度では適合文書と非適合文書が混在していて、これらの類似度を有する適合文書のみを選択することは不可能であり、例えば閾値を低く設定して、多くの適合文書を選択しようとすると、誤って選択される非適合文書が増大し、また逆に閾値を高く設定すると、誤って選択される非適合文書は減少するが、適合文書の数が減少してしまうという問題がある。

【００３５】すなわち、情報検索における検索式拡張手法をプロファイル更新に適用し、単純に類似度の閾値を設定することにより多くの適合文書を取得しようとすると、非適合文書も多く選択されてしまうという問題がある。

【００３６】本発明は、上記に鑑みてなされたもので、その目的とするところは、誤って選択された非適合文書の特徴を表す非適合プロファイルを作成し、非適合プロファイルとの類似度が高い文書を選択しないということで非適合文書の選択を回避し得る文書フィルタリング方法および装置を提供することにある。

【００３７】

【課題を解決するための手段】上記目的を達成するため、請求項１記載の本発明は、ユーザの要求を表す適合プロファイルに適合する文書を検索対象文書の中から抽出して出力する文書フィルタリング方法であって、前記適合プロファイルに対して類似度が高いが、ユーザの要求に適合しないと判定された非適合文書を特徴づけるプロファイルを非適合プロファイルとして作成し、前記適合プロファイルと検索対象文書との類似度を算出し、この算出した類似度を所定の適合用閾値と比較し、該類似度が所定の適合用閾値より大きい場合、前記検索対象文書と前記非適合プロファイルとの類似度を算出し、この非適合プロファイルとの類似度を所定の非適合用閾値と比較し、該類似度が所定の非適合用閾値より大きい場合、該検索対象文書を非適合文書と見なして除外し、小さい場合、適合文書と見なして選択し、前記非適合文書および適合文書の適合フィードバックを行い、適合プロファイルおよび非適合プロファイルを更新することを要旨とする。

【００３８】請求項１記載の本発明にあっては、適合プロファイルに対して類似度が高いが、ユーザの要求に適合しないと判定された非適合文書を特徴づけるプロファイルを非適合プロファイルとして作成し、適合プロファイルと検索対象文書との類似度が所定の適合用閾値より大きい場合、検索対象文書と非適合プロファイルとの類似度を算出し、この類似度が所定の非適合用閾値より大きい場合、検索対象文書を非適合文書と見なして除外し、小さい場合、適合文書と見なして選択し、適合フィードバックを行い、適合プロファイルおよび非適合プロファイルを更新するため、適合文書とともに選択される非適合文書の数を少なくすることができ、フィルタリング性能を向上することができる。

【００３９】また、請求項２記載の本発明は、請求項１記載の発明において、前記非適合プロファイルの更新が、選択文書に出現する単語から単語寄与度に基づき選択文書の特徴を表す単語を抽出し、この抽出された単語の寄与度を算出し、この単語寄与度に重みをかけて、単語に対するスコアを算出し、前記選択文書が適合文書である場合には、前記スコアを非適合プロファイルから減算し、前記選択文書が非適合文書である場合には、前記スコアを非適合プロファイルに加算して更新することを要旨とする。

【００４０】請求項２記載の本発明にあっては、選択文書の特徴を表す単語を抽出し、この抽出された単語の寄与度を算出し、この単語寄与度に重みをかけて、単語に対するスコアを算出し、選択文書が適合文書である場合には、スコアを非適合プロファイルから減算し、前記選択文書が非適合文書である場合には、前記スコアを非適合プロファイルに加算して更新するため、このように更新される非適合プロファイルを使用することにより文書フィルタリング精度を向上することができる。

【００４１】更に、請求項３記載の本発明は、ユーザの要求を表す適合プロファイルに適合する文書を検索対象文書の中から抽出して出力する文書フィルタリング装置であって、前記適合プロファイルに対して類似度が高いが、ユーザの要求に適合しないと判定された非適合文書を特徴づけるプロファイルを非適合プロファイルとして作成する非適合プロファイル作成手段と、前記適合プロファイルと検索対象文書との類似度を算出する適合プロファイル類似度算出手段と、この算出した類似度を所定の適合用閾値と比較し、該類似度が所定の適合用閾値より大きい場合、前記検索対象文書と前記非適合プロファイルとの類似度を算出する非適合プロファイル類似度算出手段と、この非適合プロファイルとの類似度を所定の非適合用閾値と比較し、該類似度が所定の非適合用閾値より大きい場合、該検索対象文書を非適合文書と見なして除外し、小さい場合、適合文書と見なして選択する選択手段と、前記非適合文書および適合文書の適合フィードバックを行い、適合プロファイルおよび非適合プロファイルを更新する更新手段とを有することを要旨とする。

【００４２】請求項３記載の本発明にあっては、適合プロファイルに対して類似度が高いが、ユーザの要求に適合しないと判定された非適合文書を特徴づけるプロファイルを非適合プロファイルとして作成し、適合プロファイルと検索対象文書との類似度が所定の適合用閾値より

大きい場合、検索対象文書と非適合プロファイルとの類似度を算出し、この類似度が所定の非適合用閾値より大きい場合、検索対象文書を非適合文書と見なして除外し、小さい場合、適合文書と見なして選択し、適合フィードバックを行い、適合プロファイルおよび非適合プロファイルを更新するため、適合文書とともに選択される非適合文書の数を少なくすることができ、フィルタリング性能を向上することができる。

【００４３】請求項４記載の本発明は、請求項３記載の発明において、前記更新手段が、選択文書に出現する単語から単語寄与度に基づき選択文書の特徴を表す単語を抽出する単語抽出手段と、この抽出された単語の寄与度を算出する単語寄与度算出手段と、この単語寄与度に重みをかけて、単語に対するスコアを算出するスコア算出手段と、前記選択文書が適合文書である場合には、前記スコアを非適合プロファイルから減算し、前記選択文書が非適合文書である場合には、前記スコアを非適合プロファイルに加算して更新するスコア加減算手段とを有することを要旨とする。

【００４４】請求項４記載の本発明にあっては、選択文書の特徴を表す単語を抽出し、この抽出された単語の寄与度を算出し、この単語寄与度に重みをかけて、単語に対するスコアを算出し、選択文書が適合文書である場合には、スコアを非適合プロファイルから減算し、前記選択文書が非適合文書である場合には、前記スコアを非適合プロファイルに加算して更新するため、このように更新される非適合プロファイルを使用することにより文書フィルタリング精度を向上することができる。

【００４５】

【発明の実施の形態】以下、図面を用いて本発明の実施の形態を説明する。図１は、本発明の一実施形態に係る文書フィルタリング方法を実施する文書フィルタリング装置の構成を示すブロック図である。同図に示す文書フィルタリング装置は、ユーザの要求であるプロファイルや検索対象文書などを入力する入力部１と、適合プロファイルとの類似度を算出する適合プロファイル類似度算出部３ａおよび非適合プロファイルとの類似度を算出する非適合プロファイル類似度算出部３ｂからなる適合性判定部３と、ユーザの要求に適合する文書をユーザに出力する出力部５と、プロファイルなどの情報を記憶する記憶部７と、プロファイルを更新するプロファイル更新部９と、適合フィードバックを行うフィードバック部１１とから構成されている。

【００４６】次に、上述したように構成される本実施形態の文書フィルタリング装置の作用を説明する前に、本発明の文書フィルタリング方法の概要について説明する。上述したように、従来使用されてきたプロファイル（以下、q_Rと略称する）は、ユーザの要求を表すように適合文書の特徴を取り入れて更新されてきた。しかし、q_Rに類似している文書を選択するだけでは、非適合文書も多く選択されてしまうということが明らかになっている。従って、プロファイルとの類似度が高いが実際には要求に適合していない非適合文書を選択しないようにすれば、精度を向上させることができると考えられる。そこで、過去にq_Rとの類似度が高いと判断され選択された非適合文書を特徴づけるプロファイルを非適合プロファイル（以下、q_Nと略称する）として作成する。q_Nとの類似度が高い文書は、過去に誤って選択された非適合文書集合に類似しているため、このような文書を選択しないようにすれば、従来手法では誤って選択されていた非適合文書の選択を回避することができる。

【００４７】このような考えに基づく本発明の文書フィルタリング方法について更に具体的に説明する。

【００４８】まず、従来のプロファイル更新同様、選択された文書から単語寄与度に基づき単語を抽出する。そして、抽出された単語が適合文書中の単語の場合には、単語に対するスコア$Score_{relN}(w_i)$を式（１６）により算出し、非適合文書中の単語の場合には$Score_{nrelN}(w_i)$を式（１７）によって算出する。

【００４９】

【数９】

$$Score_{relN}(w_i) = wgt_{relN} \times Cont(w_i, q, d) \quad \cdots (16)$$

$$Score_{nrelN}(w_i) = wgt_{nrelN} \times Cont(w_i, q, d) \quad \cdots (17)$$

次に、上記の式によって求めた各単語のスコアを単語出現頻度ｔｆとして扱い、以下の式（１８），（１９）により各単語w_iのＴＦ＊ＩＤＦ値を算出する。

【００５０】

【数１０】

$$Value_{relN}(w_i) = \log(1+Score_{relN}(w_i)) \times \log\left(\frac{M}{df_i}\right) \quad \cdots (18)$$

$$Value_{nrelN}(w_i) = \log(1+Score_{nrelN}(w_i)) \times \log\left(\frac{M}{df_i}\right) \quad \cdots (19)$$

ただし、df_iは単語w_iが出現する文書数、Mはdfのリスト作成に使用された文書数とする。

【００５１】また、q_Nおよび更新後のq_Nを式（２０），式（２１）で表すとする。

【００５２】

【数１１】

$$\vec{q}_N = (q_{N1}, \cdots, q_{Nn}) \quad \cdots (20)$$

$$\vec{q}_{Nnew} = (q_{N1}', \cdots, q_{Nn}') \quad \cdots (21)$$

そして、q_Rの更新とは逆に、抽出された単語w_iが適合文書中の単語の場合は、式（２２）のように、q_Nを表すベクトルから$Value_{relN}(w_i)$を引き、非適合文書中の単語の場合は、式（２３）のように、$Value_{nrelN}(w_i)$を加えることで、非適合文書の特徴を表すようにq_Nを更新する。

【００５３】

【数１２】

$$q_{Ni}' = q_{Ni} - Value_{relN}(w_i) \quad \cdots (22)$$

$$q_{Ni}' = q_{Ni} + Value_{nrelN}(w_i) \quad \cdots (23)$$

そして、q_Rとの類似度が閾値を超えた文書について、q_Nとの類似度を計算し、q_Nに対する閾値を超えた文書は過去に誤って選択した非適合文書に類似していると判断し、選択しない。

【００５４】次に、図２に示すフローチャートを参照して、図１に示す実施形態における非適合プロファイルを利用した文書フィルタリングの処理手順について説明する。

【００５５】まず、ユーザの要求である適合プロファイルq_Rと検索対象文書ｄとの類似度Sim_Rを計算し（ステップＳ１１）、この計算した類似度Sim_Rが所定の閾値$Thres_R$よりも大きいか否かを判定する（ステップＳ１３）。計算した類似度Sim_Rが所定の適合用閾値$Thres_R$より小さい場合には、最初のステップＳ１１に戻り、次の検索対象文書に対して同じ処理を繰り返し行うが、類似度Sim_Rが所定の適合用閾値$Thres_R$よりも大きい場合には、非適合プロファイルq_Nと検索対象文書ｄとの類似度Sim_Nを計算する（ステップＳ１５）。

【００５６】この計算した類似度Sim_Nが所定の非適合用閾値$Thres_N$より小さいか否かを判定する（ステップＳ１７）。類似度Sim_Nが所定の非適合用閾値$Thres_N$より小さくない場合には、すなわち類似度Sim_Nが所定の非適合用閾値$Thres_N$より大きい場合には、検索対象文書を非適合文書と見なして選択せず、すなわち除外し、最初のステップＳ１１に戻り、次の検索対象文書に対して同じ処理を繰り返し行うが、類似度Sim_Nが所定の非適合用閾値$Thres_N$よりも小さい場合には、適合文書と見なして、ユーザに出力し、これに対する適合フィードバックを得て（ステップＳ１９）、適合プロファイルq_Rおよび非適合プロファイルq_Nを上述した式（２２），（２３）に示したように更新する（ステップＳ２１）。以上の処理を全ての検索対象文書について繰り返し行う（ステップＳ２３）。

【００５７】以上の処理により適合プロファイルq_Rに類似していると判断された文書から、過去に誤って選択された非適合文書に類似している文書を削除し、選択された非適合文書の数を減少させることができる。

【００５８】次に、図３に示すフローチャートを参照して、上述した図２の処理手順において非適合文書であると見なした文書をシステムにフィードバックするpseudo feedbackによる適合プロファイルを利用した文書フィルタリング処理について説明する。

【００５９】図３に示す文書フィルタリング処理は、図

２に示した文書フィルタリング処理におけるステップＳ１７での類似度$Ｓｉｍ_N$が所定の非適合用閾値$Ｔｈｒｅｓ_N$より小さいか否かについての判定において、類似度$Ｓｉｍ_N$が所定の非適合用閾値$Ｔｈｒｅｓ_N$よりも小さくない場合に、すなわち類似度$Ｓｉｍ_N$が所定の非適合用閾値$Ｔｈｒｅｓ_N$よりも大きい場合に、検索対象文書を非適合文書と見なして、pseudo feedbackし（ステップＳ３１）、このpseudo feedbackによる情報で非適合プロファイルq_Nの更新を行うものである（ステップＳ３３）。その他の作用は図２の作用と同じであり、図３において図２と同じ処理には同じステップ番号が付されている。

【００６０】このように類似度$Ｓｉｍ_N$が所定の非適合用閾値$Ｔｈｒｅｓ_N$よりも大きいものを非適合文書と見なして、pseudo feedbackし、このpseudo feedbackによる情報も非適合プロファイルの更新に利用することで、非適合文書の更新に利用する情報を増大することができ、これにより非適合プロファイルq_Nとの類似度の閾値を厳しくした際の選択文書の減少に伴う非適合プロファイルq_Nの更新に利用するフィードバック情報の減少を補うことができる。

【００６１】すなわち、上述したように、pseudo feedbackを取り入れることにより、非適合プロファイルq_Nとの閾値を厳しくすることができ、非適合プロファイルq_Nとのフィルタリングの効果を生かし、かつ多くの文書情報を基に有効な非適合プロファイルq_Nを作成することができる。

【００６２】次に、図４に示すフローチャートを参照して、上述した図２および図３における非適合プロファイルの更新処理について説明する。

【００６３】図４に示すように、ベクトル空間モデルによって表されたプロファイルｑおよび更新後のプロファイルq_{new}、および選択文書ｄに対して、まず選択文書ｄから単語集合Ｗを抽出し（ステップＳ５３）、選択文書ｄが適合文書であるか否かを判定する（ステップＳ５７）。

【００６４】そして、選択文書ｄが適合文書である場合には、各単語に対するスコアを式（１６）のように算出し（ステップＳ５８）、それから式（２２）で示したと同様に各単語w_iのスコアをプロファイルｑから引く（ステップＳ６１）。また、選択文書ｄが適合文書でない場合には、各単語に対するスコアを式（１７）のように算出し（ステップＳ５７）、式（２３）で示したと同様に各単語w_iのスコアをプロファイルｑに加え（ステップＳ６３）、このような加減算したプロファイルｑを更新後のプロファイルq_{new}として設定する（ステップＳ６５）。

【００６５】

【発明の効果】以上説明したように、本発明によれば、ユーザの要求に適合しないと判定された非適合文書を特徴づけるプロファイルを非適合プロファイルとして作成し、適合プロファイルとの類似度が大きい場合、非適合プロファイルとの類似度を算出し、この類似度が大きい場合、検索対象文書を非適合文書と見なして除外するので、適合文書とともに選択される非適合文書の数を少なくすることができ、フィルタリング性能を向上することができる。

【００６６】また、本発明によれば、選択文書の特徴を表す単語を抽出し、この抽出された単語の寄与度を算出し、単語寄与度に重みをかけてスコアを算出し、選択文書が適合文書である場合には、スコアを非適合プロファイルから減算し、選択文書が非適合文書である場合には、スコアを非適合プロファイルに加算して更新するので、非適合プロファイルを適確に更新して文書フィルタリング精度を向上することができる。

【図面の簡単な説明】

【図１】本発明の一実施形態に係る文書フィルタリング方法を実施する文書フィルタリング装置の構成を示すブロック図である。

【図２】図１に示す実施形態における非適合プロファイルを利用した文書フィルタリングの処理手順を示すフローチャートである。

【図３】図２に示した文書フィルタリング処理において非適合文書であると見なした文書をシステムにフィードバックするpseudo feedbackによる適合プロファイルを利用した文書フィルタリング処理を示すフローチャートである。

【図４】図１の実施形態における非適合プロファイルの更新処理を示すフローチャートである。

【図５】プロファイル更新を利用した従来の文書フィルタリング処理を示すフローチャートである。

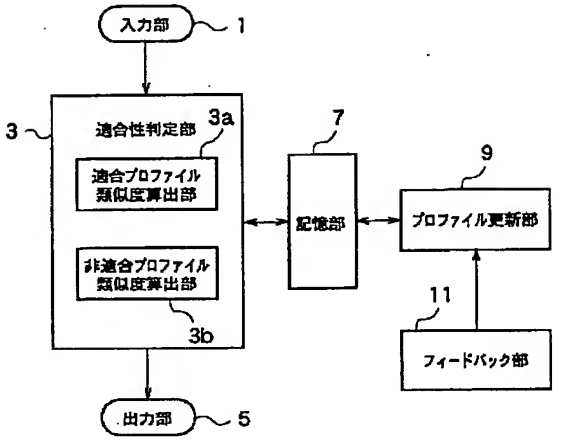
【図６】図５に示した文書フィルタリング処理におけるプロファイル更新処理を示すフローチャートである。

【図７】図５に示した従来の文書フィルタリング処理を評価する実験結果である適合文書と非適合文書の類似度を示すグラフである。

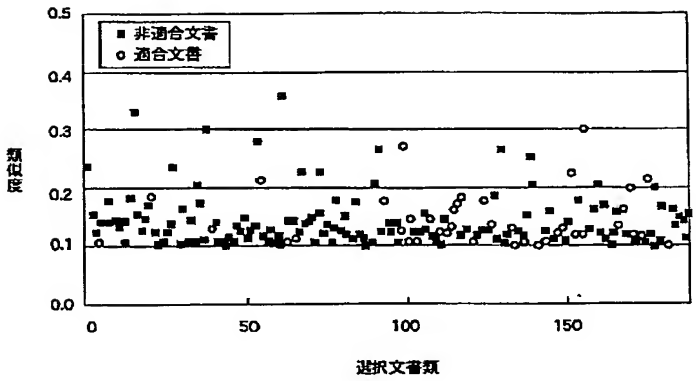
【符号の説明】

１　入力部
３　適合性判定部
３ａ　適合プロファイル類似度算出部
３ｂ　非適合プロファイル類似度算出部
５　出力部
７　記憶部
９　プロファイル更新部
１１　フィードバック部

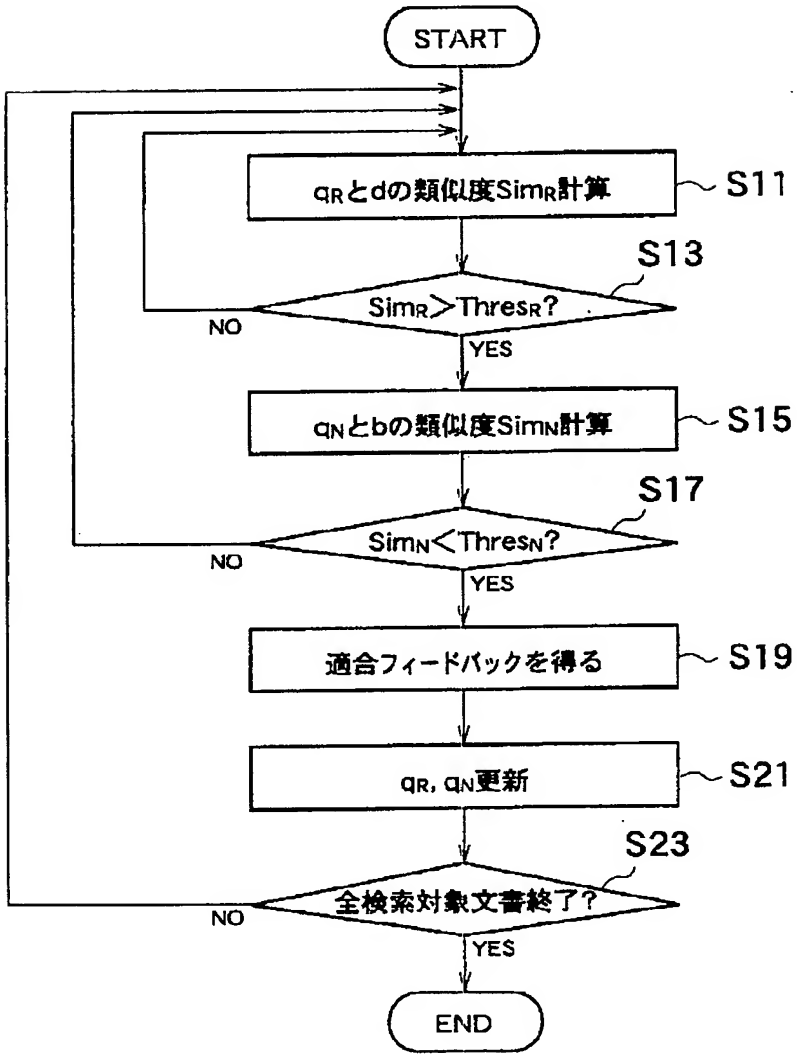
【図１】


入力部
1
3
適合性判定部
3a
適合プロファイル類似度算出部
非適合プロファイル類似度算出部
3b
7
記憶部
9
プロファイル更新部
11
フィードバック部
出力部
5

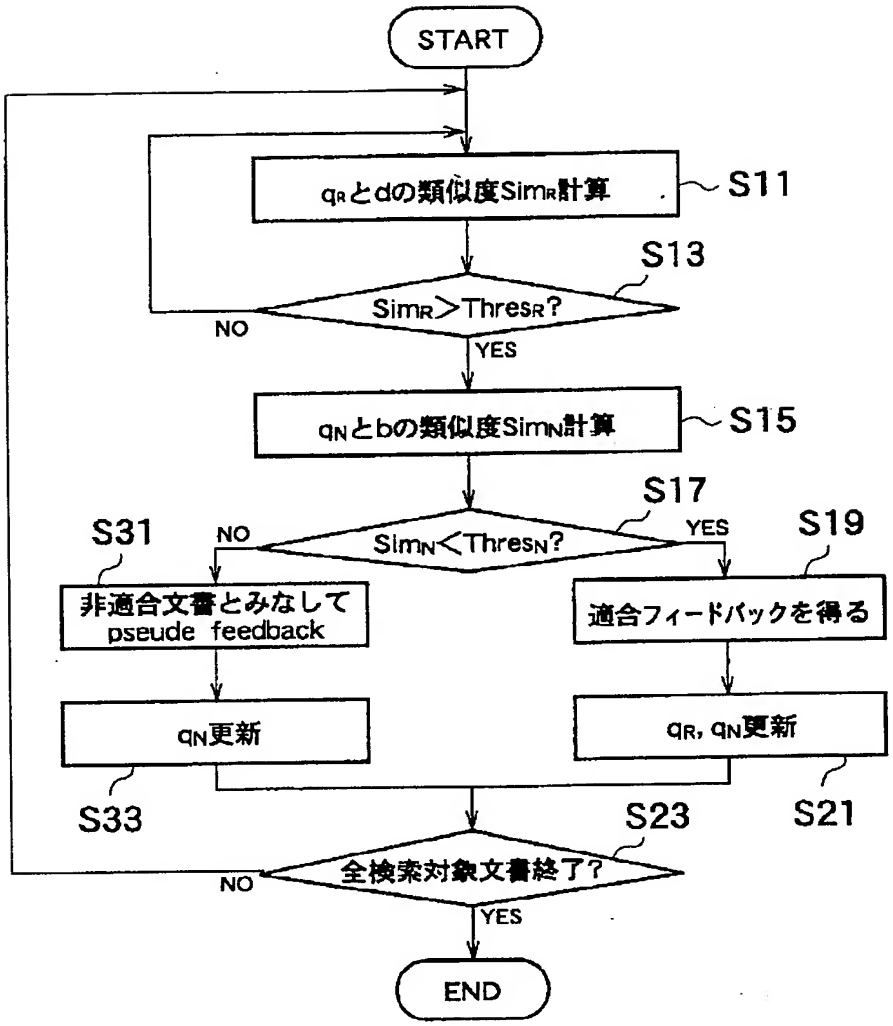
【図７】


非適合文書
適合文書
類似度
0.5
0.4
0.3
0.2
0.1
0.0
0
50
100
150
選択文書数

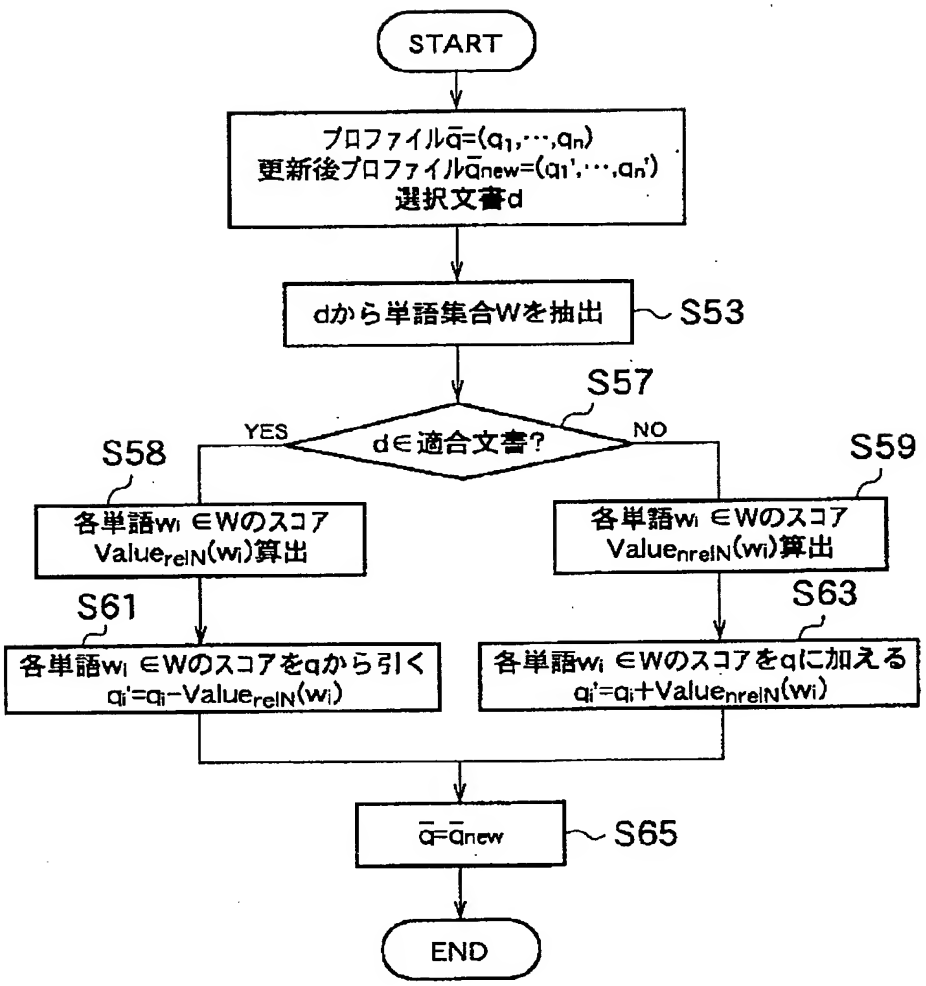
【図２】


START
qRとdの類似度SimR計算
S11
S13
SimR>ThresR?
NO
YES
qNとbの類似度SimN計算
S15
S17
SimN<ThresN?
NO
YES
適合フィードバックを得る
S19
qR, qN更新
S21
S23
全検索対象文書終了?
NO
YES
END

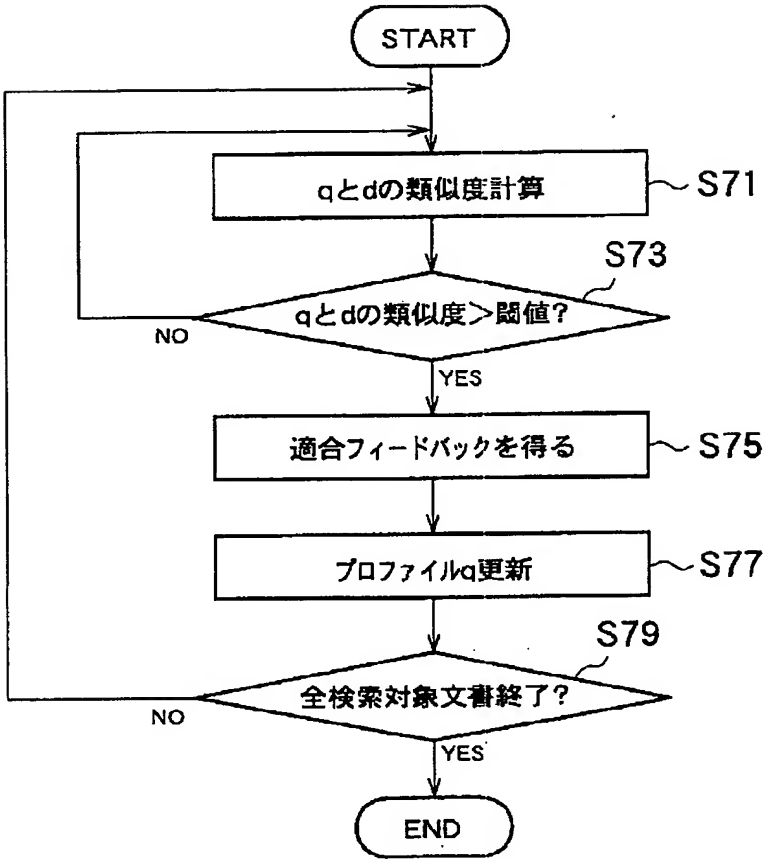
【図３】


START
q_Rとdの類似度Sim_R計算
S11
S13
$Sim_R > Thres_R$?
NO
YES
q_Nとbの類似度Sim_N計算
S15
S17
$Sim_N < Thres_N$?
NO
YES
S31
S19
非適合文書とみなして
pseude feedback
適合フィードバックを得る
q_N更新
q_R, q_N更新
S33
S23
S21
全検索対象文書終了?
NO
YES
END

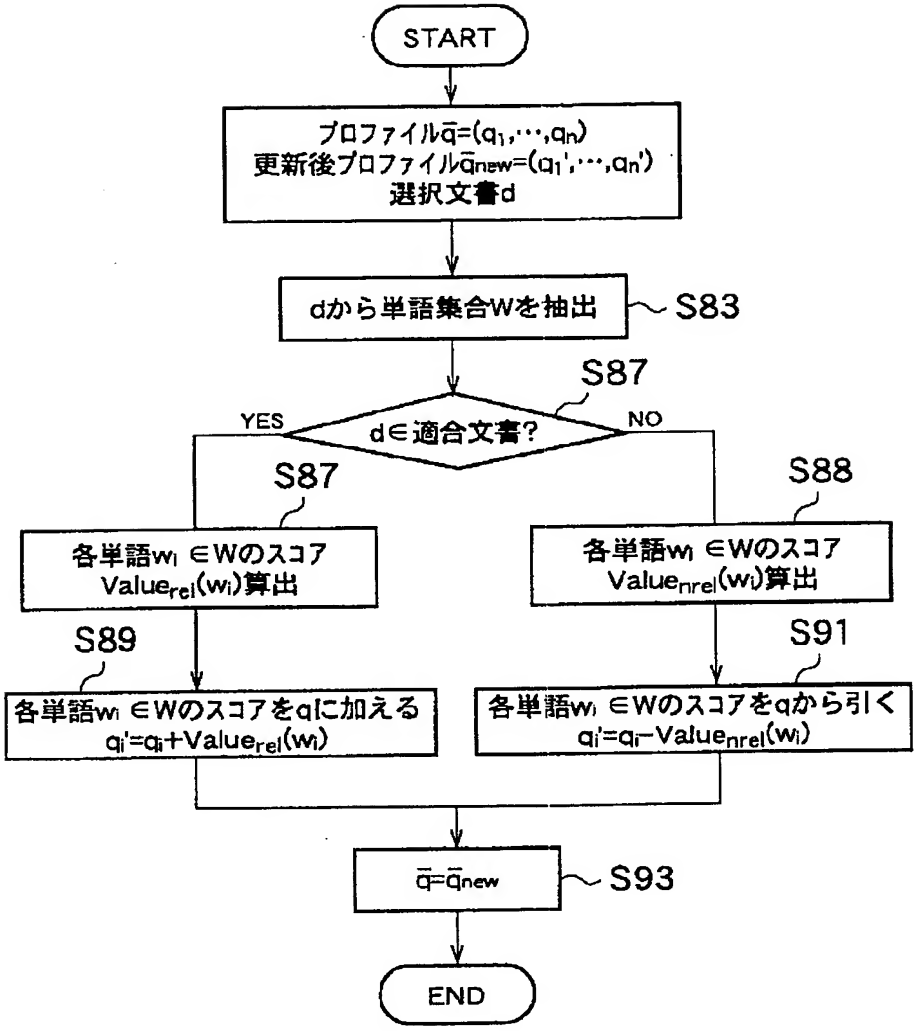
【図４】


START
プロファイル$\bar{q}=(q_1,\cdots,q_n)$
更新後プロファイル$\bar{q}_{new}=(q_1',\cdots,q_n')$
選択文書d
dから単語集合Wを抽出
S53
S57
d∈適合文書?
YES
NO
S58
各単語w_i ∈Wのスコア
$Value_{relN}(w_i)$算出
S59
各単語w_i ∈Wのスコア
$Value_{nrelN}(w_i)$算出
S61
各単語w_i ∈Wのスコアをqから引く
$q_i'=q_i-Value_{relN}(w_i)$
S63
各単語w_i ∈Wのスコアをqに加える
$q_i'=q_i+Value_{nrelN}(w_i)$
$\bar{q}=\bar{q}_{new}$
S65
END

【図５】


START
qとdの類似度計算
S71
S73
qとdの類似度＞閾値？
NO
YES
適合フィードバックを得る
S75
プロファイルq更新
S77
S79
全検索対象文書終了？
NO
YES
END

【図６】



フロントページの続き

(72)発明者　松本　一則
埼玉県上福岡市大原２－１－15　株式会社ケイディディ研究所内

(72)発明者　橋本　和夫
埼玉県上福岡市大原２－１－15　株式会社ケイディディ研究所内

Ｆターム(参考)　5B075 ND03 PR06 PR08 QM08